The Diseases Clustering for Multi-Source Medical Sets

Liangchi Li [1], Shuaijing Xu [1], Shenling Wang [1], Xianlin Ma[2*]
[1] College of Information Science and Technology, Beijing Normal University, Beijing, 100875, P.R. China
[2] Beijing Rehabilitation Hospital of Capital Medical University, Beijing, 100144, P.R. China

Abstract—**The construction of medical database has been constructed to some degrees, but the data redundancy between many medical sets has great influence on searching cross different sets. In this paper, the first step is to use three major domestic medical sets as the foundation of the research. And the Natural Language processing technologies is applied to realize the segmentation of disease description. Then, we use TF-IDF to calculate the weight of the feature words in the disease description, and establish the disease feature vector. Based on this vector, the similarity of disease feature vectors is measured by the cosine similarity method. Finally, the effect of k-means and k-center clustering algorithm on the alignment of the disease text is compared. The experimental results show that the k-center clustering algorithm has better performance compared to k-means. And the result of the clustering is reasonable to some extent.**

Keywords- Multi-source medical sets; Align diseases; text clustering; Natural Language Processing

I. INTRODUCTION

Comprehensive Chinese medical data set is the foundation of medical data mining. Due to the description from different sources of Chinese medical data set is quite different for the same disease, the fusion of data set to become the focus of the construction of Chinese medical data set.

Clustering the different disease entity is the main idea of the fusion of data sets. Since the disease descriptions are Chinese, we can't deal with them directly. In addition, the selection of cluster algorithm has influence on disease clustering.

An approach to solve the problem is proposed as follows. We deliver related literature reviews and the procedure in Section II, and describe disease data acquisition and vectorization in section III. We propose the clustering method of the disease sets in section IV and present the experimental results and their analysis as effect verification of the approach in section V. We finally account conclusion in section VI.

II. RELATED WORKS

Clustering analysis is a method to organize similar data into different groups and has become an important research field of the data mining [1-3]. In order to achieve the fusion data sets, foreign researcher established Ontology Mapping database, such as Lexical Owl Ontology Matcher (LOOM) [4,5], Unified Medical Language System Concept Unique Identifier (UMLS_CUI) etc. The mapping relationship is established by the fuzzy matching method of the disease name, but this simple relationship is greatly limited by the definition of disease name [6]. The document [7] mentioned another method to build ontology mapping, first by using the existing Bio-Portal ontology [8] mapping relation to build the similar cluster, and then using the mappings between clusters to establish ontology mapping relations, the mapping relationship between each ontology finally dug up hidden. While this approach greatly enriched the network mapping, but it is not set up the mappings between the same descriptions and different name of diseases. The semantic model for medical is proposed in [9], but they build the mapping by doctors, experts. In this paper, we propose a method based on semantic relation to realize the fusion of multi-source data set. In our scheme, the disease descriptions from different sets are firstly vectorized, thus the Chinese description is converted into Mathematical description. Then we use the partition clustering algorithm to cluster the disease vectors, and finally the relation between the different names but same descriptions of diseases can be built by the semantic relations.

A. The source of the medical data

In view of the lack of domestic medical sets [10], in this paper, the data is based on the public, the government and the business of the three major medical data sets, namely, China's public health network, baikemingyi, feihua sets. Three data sets have abundant disease information, but they have different focus. China's public health network based on civilian medical, and its collection and recording are common diseases. Baikemingyi are written by scholars and researchers, so the description of the disease and introduce mainly based on the view of the academic. The data of feihua sets are from the clinical disease and the corresponding symptoms. In order to more fully reflect the characteristics of the disease, this article take the method mentioned before to build a more complete set of medical data.

978-1-5090-5952-2/16 $31.00 © 2016 IEEE
DOI 10.1109/IIKI.2016.37

B. *Clustering analysis*

The choice of clustering methods for disease entity is also the focus of the research. Partitioning methods are widely used in document clustering for its low complexity. The most popular and simple partitioning clustering algorithms, k-means, is still widely used in document clustering. The partition clustering algorithm in data mining is given in [11], the research about the k-means algorithm applied in the Chinese document is presented in [12]. However, since the k-means is sensitive to the noisy, we take the k-center to realize the clustering. For better clustering the different disease description, the measurement becomes the importance of the research. Many methods about the measurement is proposed in [13]. In this paper, we combine the k-center algorithm and take the cosine similarity as the measure of the disease vector.

III. DATA ACQUISITION AND VECTORIZATION

The three Chinese data sets mentioned before can't be directly obtained through the API interface, these are often embedded in complex web structure. In order to extract the disease information, this paper mainly take the web crawler technology [14].

The web crawler technology is mainly based on the following steps. Firstly, the medical website is set as the initial URL of the crawler, and then the python scripts camouflage browser sends a request to the server. From analyzing the response of the server, we can extract the disease description and generate a new URL, and only when the scripts grab and analyze all pages is the cycle come to the end. The Fig.1 show the detail process of the data acquisition.

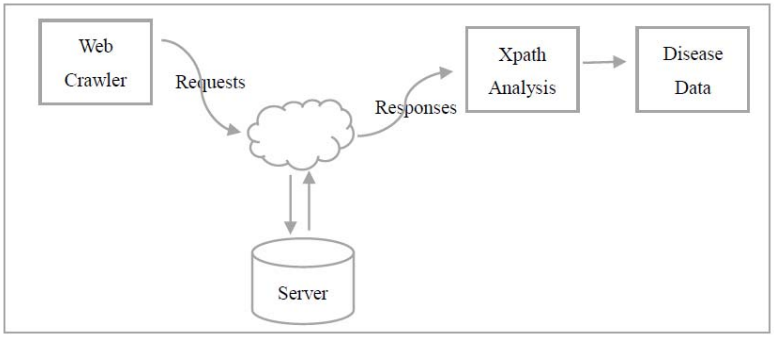

Fig. 3. Data acquisition in the web crawler

How to transfer the unstructured disease description into structured vector model become another focus of this paper. The most widely used method of representation is the vector space model (VSM). The basic idea of the VSM is to transform disease description into vector.

This paper uses the word character of the Term Frequency (TF) and Inverse Document Frequency (IDF) to vectorize the disease description. First, we adopt the Natural Language Processing technologies based on the Chinese academy of sciences word Segmentation technology [15] to transform the disease description into many simple expressions, and remove the stop words. Then we extract the rest of the word as the characterizations of disease description. In the following steps, we merely extract the key words in one of the disease description. The key word vector must contain all the key words in different description. Specially, since the disease name is of great importance to the disease description, the disease name is also added to the key word vector. We assume that the key word vector is defined as follows:

$$\boldsymbol{V} = \{k_1, k_2, \dots, k_N, DN_1, DN_2, \dots, DN_M\} \tag{1}$$

Where the k_i is the i^{th} key words and the N represent the number of the key words. The DN_i is the i^{th} disease name. The M represent the number of the disease. The word vector is (N+M)-dimensional. So the disease description can be defined as:

$$\boldsymbol{D_j} = \{\omega_{k_1}, \dots, \omega_{k_N}, \omega_{DN_1} \dots, \omega_{DN_M}\}, \ \ 1 \le j \le M + N \tag{2}$$

Where the D_i is the disease vector. And finally, by calculating the Term Frequency (TF) and Inverse Document Frequency (IDF) [16] to vectorize the disease description. The disease weight matrix can be written as:

$$\omega_{k_i} = \begin{cases} 0, & if \ \ k_i \notin \boldsymbol{V} \\ TF\text{-}IDF(k_i), & else \end{cases} \tag{3}$$

Where the ω_{k_1} is the weight of the word k_i. The Di is the i^{th} disease description. The $TF - IDF(k_i)$ is the term k_i in the document matrix D_j.

IV. CLUSTERING OF THE DISEASE SETS

The selection of clustering algorithm is the key to disease clustering. Due to any medical sets can be broadly include all diseases, so the initial number of the clusters can be set as the disease number in any set. Based on clustering cluster number known as the prior knowledge, the clustering algorithm based on partition method. In this paper, we choice the minimum number of data sets as the cluster number to keep the differences and similarity between the clusters of disease.

The appropriate partition method directly influences the effect of the clustering. Considering the partition method, we often take the k-means and k-center algorithm into account. Compared to the k-means clustering algorithm, the k center clustering algorithm has better noise immunity performance. The k-center clustering algorithm firstly partition each entity to the closest cluster, then set the nearest entity as the new center of the cluster, and iterative process above, until the cluster of each entity is not changed.

Algorithm 1

Input: 1) the disease matrix 2) the number of initial clusters

Output: the k clusters

1. Randomly select k vector as the initial cluster center

2. **Repeat**

3. According to the similarity measure the k clusters, each disease is assigned to the most similar cluster.

4. Calculate the most similarity disease vector as the new cluster center.

5. Until all the disease vector no longer change their cluster

The choice of measurement is the focus of the research. The similarity measurement depends on the different practical problems. Common partitioning clustering uses the Euclidean distance as the similarity measurement method, but in the medical environment, due to illness profiles feature vector is too sparse, we use cosine similarity method. The greater the vector cosine value is, the more similar the vector is, that is, the closer the two disease situation.

V. SIMULATION AND RESULTS

The data acquisition of the web crawler technology is not only accurate, but also fast. Table 1 show the number of diseases of different sets. The diseases in *China's public health network* are summarized of the Common diseases, so the number of it may be smaller than that of other sets. Since the data of the *feihua sets* is from the clinical disease, the range of the disease is larger than others. From the analyzed result, the total number of three sets is 18,675.

Table 1. Number of different of data sets

Source of the data	Diseases number
China's public health network	3633
Baikemingyi	5224
feihua sets	9818

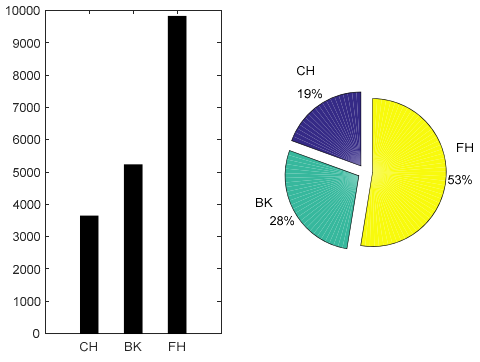

Fig.2. Comparison between the medical sets

The left picture in Fig.2 shows the quantitative relation between three sets. The right picture in Figure.2 shows each data set in the proportion of all data.

In this section, we present some simulation result of the cluster. After the vectorization of the disease description, 18,675 disease descriptions are transformed into mathematics vector. Then we may select proper algorithm to cluster the

vector, who has the nearest measurement. Based on the partitioning methods, we naturally took the k-means algorithm and k-center algorithm into account. The Fig.3 show the comparison of the consequence between the two algorithms.

Table 2. Result of the Disease Cluster

	k-means	k-center
Greater than 2 entity	12479(86.49%)	13296(92.15%)
Majority cluster size	3	3

Table 3. Four conditions of the Disease

	Name	Description
CLASS-1	Same	Same
CLASS-2	Same	Different
CLASS-3	Different	Same
CLASS-4	Different	Different

According to the Table 2, it can be seen the different result of k-means and k-center algorithm in clustering. After the fusion of the three medical sets, the same disease from different sets can be form a new cluster whose size may be three. So just considering the size of majority cluster, both methods have good performance in clustering. However, the k-center algorithm may be better than k-means as shown in Fig.3. The size of most clusters are greater than 2 in the k-center algorithm, but the result of k-means has many one or two clusters.

Compared two kinds of diseases in the cluster, we may distinguish it by its name and description. In this paper, we divide the way distinguish it into four conditions. The detail of the condition is listed in the Table 3.

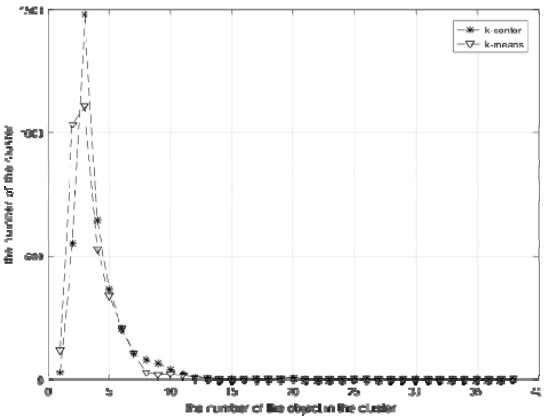

Fig.3. the cluster result of k-means and k-center

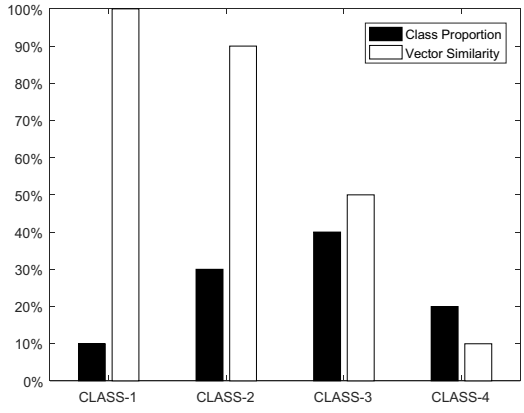

Fig.4. Class Proportion and Vector Similarity

From analyzing the disease in the cluster, the relationship between the two kinds of diseases under the same cluster also accord with the laws of the table. Through the statistical analysis of disease under all clusters, the disease relations and corresponding similarity is roughly three results conform to the Fig.4.

Fig.4 shows the proportion of listed CLASS in Table 3. The black block shows the proportion of the class, and the white shows the similarity of the corresponding class. Since the similarity of CLASS-1 was 100%, there is no doubt the CLASS-1 get into a cluster. As for the disease in the CLASS-2, although the description is not exactly the same, the same name of the disease will determine it. Then, in the CLASS-3, the key words in the disease description have played an important role in gathering the same disease. Finally, to the disease in CLASS-4, although the disease name and description is different, but cluster algorithm can still find an optimal clustering. The optimal clustering is accord with the facts of a cluster. For example, the name of hypertensive intracerebral hemorrhage and hypertension disease is not identical, but still have many same words, such as hypertension. The key words not only increase the similarity of the disease, but also determine the disease cluster. So when there is no identical match, intracranial hemorrhage and hypertension is also a kind of reasonable matching.

The k-center algorithm will cluster about 16942 disease entity vector. In the Table 4, the disease was gathered into 3633 clusters and the number of biggest entities in the cluster is 76. In this paper, we display the number of entity is 5 cluster in Fig.5.

Table 4. Result of the cluster

Number of entities	16942
Number of clusters	3633
Largest cluster	76
Majority cluster size	3
3 entity clusters	1481

Table 5. Cluster of the Emesis

Source	Disease Name	Similarity
FH	**Nervous Vomiting**	0.143531
	Functional Vomiting	0.169665
BK	**Emesis**	0.974672
	Nervous Vomiting	0.115499
CH	**Emesis**	1.000000

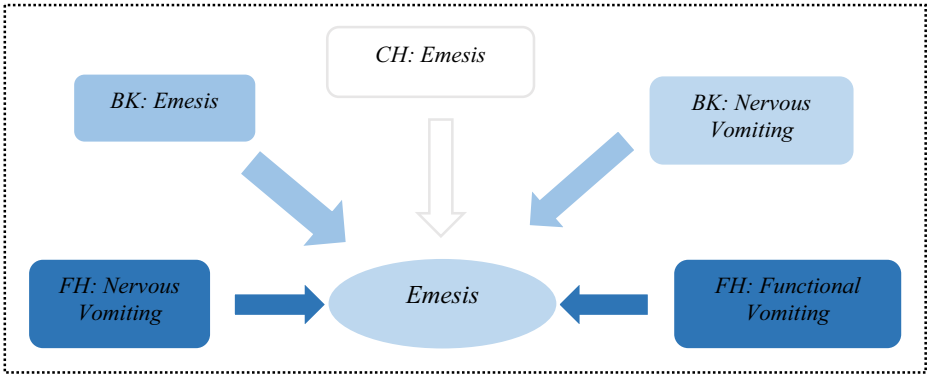

Fig.5. Cluster of the Emesis

The Fig.5 and the Table 5 show the number of entity is 5 cluster. In the Fig.5, the source of the data is in the front of each box, the back of it stand for the name of disease. From analyzing the disease, emesis, the key word of the disease, play an important role in clustering. Though nervous vomiting and functional vomiting can't match emesis exactly, the two disease which relied on the keywords of disease description had high disease similarity in Table 5. The experimental results further verify the rationality of clustering.

VI. CONCLUSION

In this paper, we succeeded in clustering the diseases under multi-source medical sets. From using the Natural Language Processing technologies and k-center algorithm, all the disease in different sets was clustered into 3,633 clusters. And the diseases within the same cluster were described in many aspects, such as the cosine similarity, the proportion of the disease and so on. According to different clustering results, we made a simple analysis.

ACKNOWLEDGMENTS

This research is sponsored by National Natural Science Foundation of China (No.61371185, 61401029, 61472044, 61472403, 61571049 61601033) and the Fundamental Research Funds for the Central Universities (No. 2014KJJCB32, 2013NT57) and by SRF for ROCS, SEM. and China Postdoctoral Science Foundation Funded Project (No.2016M590337).

REFERENCES

[1] Mehmood R, Zhang G, Bie R, Dawood H , Clustering by fast search and find of density peaks via heat diffusion, Neurocomputing (2016).

[2] R, Bie R, Jiao L, Dawood H, Sun Y. Adaptive cutoff distance: Clustering by fast search and find of density peaks. Journal of Intelligent and Fuzzy Systems. 2016 Jan 1; 31(5):2619-28.

[3] Bie R, Mehmood R. Adaptive fuzzy clustering by fast search and find of density peaks. Personal and Ubiquitous Computing. 2016 Oct 1; 20(5):785-93.

[4] Ghazvinian, A, N. F. Noy, and M. A. Musen. "Creating mappings for ontologies in biomedicine: simple methods work." AMIA. Annual Symposium proceedings / AMIA Symposium. AMIA Symposium 2009(2008):198-202.

[5] Mohammed O, Benlamri R, Fong S. Building a Diseases Symptoms Ontology for Medical Diagnosis: An Integrative Approach[C]. IEEE International Conference on Future Generation Communication Technology. 2012:104-108.

[6] Shvaiko, P., and J. Euzenat. "Ontology Matching: State of the Art and Future Challenges." IEEE Transactions on Knowledge & Data Engineering 25.1(2013):158-176.

[7] Oberkampf, Heiner, et al. "From Symptoms to Diseases - Creating the Missing Link." Extended Semantic Web Conferenece 2015:652-667.

[8] Whetzel PL; Noy NF. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications." Nucleic Acids Research 39.Web Server issue (2011):W541-5.

[9] Mohammed, O, and R. Benlamri. "Developing a semantic web model for medical differential diagnosis recommendation." Journal of Medical Systems 38.10(2014):1-18.

[10] Zhen Zhang etc. "medical data and its facing the opportunities and challenges." journal of medical informatics (2014): 35.6 2-8.

[11] Zonghu Wang. The key technology research on the optimization of the clustering analysis. Xidian University university, 2012.

[12] Suhui Wu. "The text representation and similarity computing research in the clustering." Intelligence science, 4 (2012).

[13] Tiantian Zhu. Research on the similarity measurement method and application of the short text semantic. East China normal university, 2014.

[14] Feng Qiao. Based on Web crawler technology of Web page information extraction. University of electronic science and technology, 2012.

[15] Zimu Wang. "A method of using TF - IDF combined with the text similarity measurement method research of lexical semantic information." Jilin University, 2015.

[16] Zhipeng Cai, Tong Zhang, and Xiufeng Wan. A Computational Framework for Influenza Antigenic Cartography. PLoS Computational Biology. 6 (10) (2010), e1000949.

[17] Zhipeng Cai, Mariette F. Ducatez, Jialiang Yang, Tong Zhang, Li-Ping Long, Adrianus C. Boon, Richard J. Webby, and Xiu- Feng Wan. Identifying Antigenicity Associated Sites in Highly Pathogenic H5N1 Influenza Virus Hemagglutinin by using Sparse Learning. Journal of Molecular Biology 422(1):145-155 (2012).

[18] Zhipeng Cai, Tong Zhang. Antigenic Distance Measurements for Seasonal Influenza Vaccine Selection. Vaccine. 30(2):448-453 (2012).

[19] Zhipeng Cai, Randy Goebel, Mohammad Salavatipour, and Guohui Lin. Selecting dissimilar genes for multi-class classification, an application in cancer subtyping. BMC Bioinformatics. 8(2007), 206.